

# Taming outliers in pulsar-timing data sets with hierarchical likelihoods and Hamiltonian sampling

Michele Vallisneri<sup>1,2★</sup> and Rutger van Haasteren<sup>1,2</sup>

<sup>1</sup>*Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA*

<sup>2</sup>*TAPIR, California Institute of Technology, Pasadena, CA 91125, USA*

Accepted 2017 January 10. Received 2016 December 17; in original form 2016 September 19

## ABSTRACT

Pulsar-timing data sets have been analysed with great success using probabilistic treatments based on Gaussian distributions, with applications ranging from studies of neutron-star structure to tests of general relativity and searches for nanosecond gravitational waves. As for other applications of Gaussian distributions, *outliers* in timing measurements pose a significant challenge to statistical inference, since they can bias the estimation of timing and noise parameters, and affect reported parameter uncertainties. We describe and demonstrate a practical end-to-end approach to perform Bayesian inference of timing and noise parameters *robustly* in the presence of outliers, and to identify these probabilistically. The method is fully consistent (i.e. outlier-ness probabilities vary in tune with the posterior distributions of the timing and noise parameters), and it relies on the efficient sampling of the hierarchical form of the pulsar-timing likelihood. Such sampling has recently become possible with a ‘no-U-turn’ Hamiltonian sampler coupled to a highly customized reparametrization of the likelihood; this code is described elsewhere, but it is already available online. We recommend our method as a standard step in the preparation of pulsar-timing-array data sets: even if statistical inference is not affected, follow-up studies of outlier candidates can reveal unseen problems in radio observations and timing measurements; furthermore, confidence in the results of gravitational-wave searches will only benefit from stringent statistical evidence that data sets are clean and outlier-free.

**Key words:** gravitational waves – methods: data analysis – pulsars: general.

## 1 INTRODUCTION

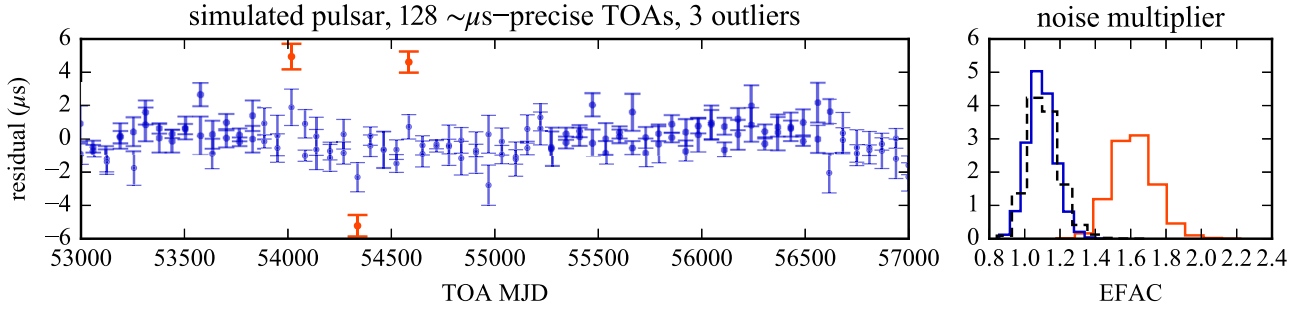
The scientific value of pulsar timing (Lorimer & Kramer 2012) lies in the possibility of performing very accurate fits of very detailed physical models, allowing remarkable applications and discoveries, such as characterizing the structure and physics of pulsars (Lattimer & Prakash 2007), testing general relativity (Stairs 2003), identifying extrasolar planets (Wolszczan & Frail 1992), mapping free-electron density across the Galaxy (Cordes & Lazio 2002, 2003) and searching for nanoHertz-band gravitational waves. See Lommen (2015) and Burke-Spolaor (2015) for recent reviews. Here we use ‘fit’ as a loose term for ‘statistical inference’ (whether of the Bayesian or classical variety), whereby a probabilistic model of the noise is used with the data to derive estimates for the timing parameters of the pulsar. The noise model may incorporate components due to the timing measurements, to intrinsic irregularities in the periodic emission of pulses and to delays induced through propagation in the interstellar medium. The timing parameters comprise of a basic rotation model, astrometric parameters and orbital elements for binary pulsars (Edwards, Hobbs & Manchester 2006; Lommen & Demorest 2013). We may even include deterministically mod-

elled or noise-like gravitational waves, and endeavour to establish their presence or to limit their amplitude, alongside our estimation of timing parameters (most recently EPTA 2015; PPTA 2015; NANOGrav 2016).

Mathematically, probabilistic models of noise are almost always based on Gaussian distributions (however, see Lentati, Hobson & Alexander 2014). The ‘radiometer’ errors incurred in measuring individual pulse times-of-arrival (TOAs) are taken as independent normal variables, each with a different variance, a function of the signal-to-noise ratio (SNR) of each observation. Even time-correlated pulsar-spin noise can be described as a Gaussian process (see van Haasteren & Vallisneri 2014 for a recent review). This leads to a likelihood – the probability of obtaining the observed data as a function of the timing-model and noise parameters – in the form of a joint Gaussian distribution. For such a likelihood, the inference problem can be solved analytically if the noise parameters are fixed and the effect of the timing parameters is linearized, or at least the problem can be attacked numerically with surprising efficiency (van Haasteren & Vallisneri 2014).

The problem with outliers: unfortunately, Gaussian likelihoods are very vulnerable to the presence of *outliers* in the data set. These are data points that have a physical origin other than the process reflected by our deterministic/probabilistic model of the data. For

\* E-mail: Michele.Vallisneri@jpl.nasa.gov



**Figure 1.** Left-hand panel: simulated pulsar-timing data set with the addition of three strong outliers (thick red markers). We show residuals computed against the best-fitting timing-model and noise parameters. Right-hand panel: posterior probability distribution for the EFAC noise multiplier, as computed in the presence of outliers (red histogram, displaced to  $\text{EFAC} \simeq 1.6$ ), after excluding them outright (blue histogram), and with the outlier-robust analysis described later in this paper (dashed black histogram). The analysis identifies all three outliers correctly, with  $P_{i,\text{out}} \simeq 1$ . The data set was generated using the `libstempo` PYTHON interface to `TEMPO2` ([github.com/vallis/libstempo](https://github.com/vallis/libstempo)) and the `libstempo/toasim` module, on the basis of the PSR J0030+0451 timing parameters (with simplifications).

instance, outliers may arise in low-SNR timing observations from strong thermal-noise spikes being mistaken for actual radio pulses; in this case, not only are the outliers spread much more broadly than ‘good’ measurements, but they do not even centre around the true TOAs. More generally, the statistical distribution of outliers can be very different (biased, much broader and non-Gaussian) than represented in our formulas, affecting the accuracy of statistical inference to a degree that depends on the number and severity of the outliers. Consider for instance the effect of outliers on a least-squares fit under the assumption of independently distributed errors: since the outliers are displaced by several standard deviations from the best fit that could be derived if the outliers were not in the data, they disproportionately affect the chi square (a quadratic function of the uncertainty-normalized deviation of data from the model) and may end up dominating the estimates of model parameters.

To make this discussion more concrete for the case of pulsar timing, in Fig. 1 we show a simulated data set of 128 TOAs (based on PSR J0030+0451, with significant simplifications in the timing model). We have introduced three large outliers, identified by the thick red dots and errorbars in the left-hand panel of the figure. The outliers bias the estimation of the TOA measurement noise: as shown by the red profile in the right-hand panel, the Bayesian posterior probability for the ‘EFAC’ parameter (which multiplies individual measurement errors in the data set) is displaced to 1.6 times the correct value of 1, to account for the additional outlier-induced variance. By contrast, the blue profile in the right-hand panel shows the estimate of noise that would be obtained if the outliers were not present. In this fit, the effect of the outliers on the timing-model parameters is only to increase their uncertainty (since the fit prefers more measurement noise) rather than to bias their estimates – which can nevertheless happen, depending on the configuration of the outliers.

## 2 OUTLIER MITIGATION

What to do? For data contamination as blatant as in Fig. 1, we may just identify the outliers visually and exclude them, or at least inspect the original TOA measurements and look for anomalies. However, such a manual solution is incompatible with reproducibility and unbiasedness, and it is also impractical for large amounts of data. A variety of more objective outlier-mitigation algorithms have been proposed in the statistical literature (Leroy & Rousseeuw 1987; Barnett & Lewis 1994; Hawkins 2013). Perhaps the simplest approach, known as *sigma clipping* (a variant of iterative deletion, see Leroy & Rousseeuw 1987), can be formulated as follows in

the context of linear least-squares estimation. Let our problem be described by

$$y_i = \sum_{\mu=1}^P M_{i\mu} \eta_{\mu} + \epsilon_i, \quad \text{for } i = 1, \dots, N; \quad (1)$$

here the  $y_i$  are the  $N$  measurements, the  $\eta_{\mu}$  are the  $P$  parameters that we wish to estimate,  $M_{i\mu}$  is the *design matrix* (whose columns may encode, e.g. a constant, a linear trend, a quadratic) and the  $\epsilon_i$  are (unknown) measurement errors taken to be independently distributed as Gaussians,  $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$  – except that some are instead outliers drawn from a different, much broader distribution.

In sigma clipping, we first fit the model using all the data, resulting in the parameter estimates  $\eta_{\mu}^{(0)}$ ; we compute the post-fit residuals  $r_i^{(0)} = y_i - \sum_{\mu} M_{i\mu} \eta_{\mu}^{(0)}$ ; we identify the data points for which  $r_i^{(0)}/\sigma_i$  is greater than a set threshold, large enough that such an error would be very unlikely to appear in the data; we deem the worst offending point an outlier, and discard it from the data set; we fit the model again, resulting in the updated (and hopefully less biased) parameter estimates  $\eta_{\mu}^{(1)}$  and residuals  $r_i^{(1)}$ ; and we continue iteratively until no residuals are found above the sigma threshold. At every step, we remove the data point that contributes the most to the fit’s  $\chi^2$ , defined as  $\sum_i r_i^2 / \sigma_i^2$ .

Sigma clipping is straightforward and makes intuitive sense, but it does not generalize well to the pulsar-timing case, for two reasons: first, because the noise parameters enter the computation of the likelihood, there is no unique set of residuals that may be used to define outlier-ness; secondly, in the presence of red timing noise or dispersion-measure fluctuations, the stochastic components of the TOAs become correlated and the likelihood has the form  $\exp\{-\mathbf{r}^T \mathbf{C}^{-1} \mathbf{r} / 2\}$ , with  $\mathbf{C}$  a dense matrix, so the contribution of each data point to  $\chi^2$  cannot be isolated.

A Bayesian mixture model of outliers: a statistically more principled procedure (advocated by Hogg, Bovy & Lang 2010, and discussed more formally in Press 1997 and Jaynes & Bretthorst 2003) follows from recognizing that least-squares estimation is equivalent to maximizing the likelihood

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\eta}) &= \prod_i p(y_i|\boldsymbol{\eta}) \\ &= \prod_i \left[ \exp \left\{ - \left( y_i - \sum_{\mu} M_{i\mu} \eta_{\mu} \right)^2 / (2\sigma_i^2) \right\} / \sqrt{2\pi\sigma_i^2} \right] \end{aligned} \quad (2)$$

(which is indeed proportional to  $e^{-\chi^2/2}$ ), and from replacing the likelihood of each individual measurement with an expression that allows for the possibility that the measurement is an outlier:

$$p'(y_j|b_i, \boldsymbol{\eta}; \sigma_{\text{out}}) = \begin{cases} e^{-(y_i - \sum_{\mu} M_{i\mu} \eta_{\mu})^2 / (2\sigma_i^2)} / \sqrt{2\pi\sigma_i^2} & \text{for } b_i = 0, \\ e^{-y_i^2 / (2\sigma_{\text{out}}^2)} / \sqrt{2\pi\sigma_{\text{out}}^2} & \text{for } b_i = 1; \end{cases} \quad (3)$$

here the  $b_i$  are binary labels that identify each  $y_i$  as either a regular data point or an outlier, and  $\sigma_{\text{out}}$  (with  $\sigma_{\text{out}} \gg \text{every } \sigma_i$ ) represents the typical range of outlier fluctuations. (Note that we are slightly modifying Hogg et al.'s treatment by modelling outliers that are not just much noisier measurements, but measurements of *noise alone*.) This likelihood can be maximized *as is* to obtain the most likely model parameters  $\eta_{\mu}$  and outlier classifications  $b_i$ . In a Bayesian inference context, if we provide a prior for the  $b_i$ ,

$$P(b_i = 1) = P_{\text{out}}, \quad P(b_i = 0) = 1 - P_{\text{out}}, \quad (4)$$

we may also *marginalize* equation (5) with respect to the  $b_i$ , yielding a remarkably simple *mixture* expression:

$$p''(y_j|\boldsymbol{\eta}; \sigma_{\text{out}}, P_{\text{out}}) = (1 - P_{\text{out}}) \times p'(y_i|b_i = 0, \boldsymbol{\eta}) + P_{\text{out}} \times p'(y_i|b_i = 1, \boldsymbol{\eta}; \sigma_{\text{out}}), \quad (5)$$

where  $p'(y_i|b_i = 0, \boldsymbol{\eta})$  and  $p'(y_i|b_i = 1, \boldsymbol{\eta}; \sigma_{\text{out}})$  are given by the two rows of equation (3). As it is manifest in equation (5), we are not completely excluding points that are exceedingly unlikely (as in sigma clipping), but instead we allow every point to behave as a regular measurement or an outlier, according to  $P_{\text{out}}$  and to the relative weight of  $p'(y_i|b_i = 0, \boldsymbol{\eta})$  and  $p'(y_i|b_i = 1, \boldsymbol{\eta}; \sigma_{\text{out}})$ .

We can now perform statistical inference using the full-data set likelihood  $p''(\mathbf{y}|\boldsymbol{\eta}; \sigma_{\text{out}}, P_{\text{out}}) = \prod_i p''(y_i|\boldsymbol{\eta}; \sigma_{\text{out}}, P_{\text{out}})$ , gaining robustness against outliers at the cost of adding the parameters  $\sigma_{\text{out}}$  and  $P_{\text{out}}$ . (In fact, these are *hyperparameters*, since they determine the form of the likelihood for the regular parameters  $\eta_{\mu}$ .) In Bayesian inference, we can hold  $\sigma_{\text{out}}$  and  $P_{\text{out}}$  fixed to reasonable values; or, more naturally, we can assign priors to them and let the data sort them out. That is, we sample (e.g. with Markov chain Monte Carlo; Liu 2013) the model parameters  $\eta_{\mu}$  together with  $\sigma_{\text{out}}$  and  $P_{\text{out}}$ , resulting in the joint parameter-hyperparameter posterior probability  $p(\boldsymbol{\eta}; P_{\text{out}}, \sigma_{\text{out}}|\mathbf{y})$ . The marginal posterior  $p(P_{\text{out}}|\mathbf{y}) = \int p(\boldsymbol{\eta}; P_{\text{out}}, \sigma_{\text{out}}|\mathbf{y}) d\boldsymbol{\eta} d\sigma_{\text{out}}$  encodes the fraction of outliers that our scheme identifies in the data, while the probability that each individual data point  $y_i$  is an outlier is given by

$$P_{i,\text{out}} = \int \frac{P_{\text{out}} \times p'(y_i|b_i = 1, \boldsymbol{\eta}; \sigma_{\text{out}})}{(1 - P_{\text{out}}) \times p'(y_i|b_i = 0, \boldsymbol{\eta}) + P_{\text{out}} \times p'(y_i|b_i = 1, \boldsymbol{\eta}; \sigma_{\text{out}})} \times p(\boldsymbol{\eta}; P_{\text{out}}, \sigma_{\text{out}}|y_i) d\boldsymbol{\eta} dP_{\text{out}} d\sigma_{\text{out}}. \quad (6)$$

A similar mixture method is described by Kunz, Bassett & Hlozek (2007, see also Hlozek et al. 2012; Knights et al. 2013) for an astronomical parameter-estimation problem where each observation may originate from different source species, each requiring a different likelihood. A mixture likelihood is also used by Abdo et al. (2013) to describe the dual origin of gamma-ray photons from either a pulsar or a diffuse background.

Application to pulsar timing: can we apply the mixture scheme to pulsar timing? The first difficulty that we outlined above for sigma clipping was the necessity of estimating the noise (hyper-)parameters, which may affect the very notion of outlier-ness. But this is no different from what already happens for  $\sigma_{\text{out}}$  and  $P_{\text{out}}$  in the

mixture scheme, so we can just sample the noise hyperparameters alongside the other two.

The second difficulty was the requirement of a likelihood that can be factorized into sublikelihoods for each individual point, whereas the most general timing-model likelihood involves a dense vector-matrix products of residuals and noise covariance. There is in fact a form of timing-model likelihood, known as *hierarchical*, which is manifestly factorizable. To explain how it comes about, we begin with the more usual time-correlation form of the likelihood:

$$p_{\text{GP}}(\mathbf{y}|\boldsymbol{\eta}) = \frac{e^{-\frac{1}{2} \sum_{ij} (y_i - \sum_{\mu} M_{i\mu} \eta_{\mu})(N_{ij} + K_{ij})^{-1} (y_j - \sum_{\nu} M_{j\nu} \eta_{\nu})}}{\sqrt{(2\pi)^n \det(N + K)}}, \quad (7)$$

where ‘GP’ stands for ‘Gaussian-process’. In this equation, the  $y_i$  are the  $n$  pulsar-timing residuals, the  $\eta_{\mu}$  are the timing parameters,  $M_{i\mu}$  is the design matrix that encodes the effect of changing the timing parameters around their best-fitting values, the diagonal matrix  $N_{ij} = \delta_{ij} \sigma_i^2$  collects the individual variance of the measurement errors (which are analogue to the  $\epsilon_i$  of equation 1) and the *dense* matrix  $K_{ij}$  represents the covariance of correlated noise, a function of a set of noise hyperparameters not shown here to simplify notation. (See van Haasteren & Vallisneri 2014 for a review of this formalism.) Because of  $K_{ij}$ , the likelihood cannot be factorized.

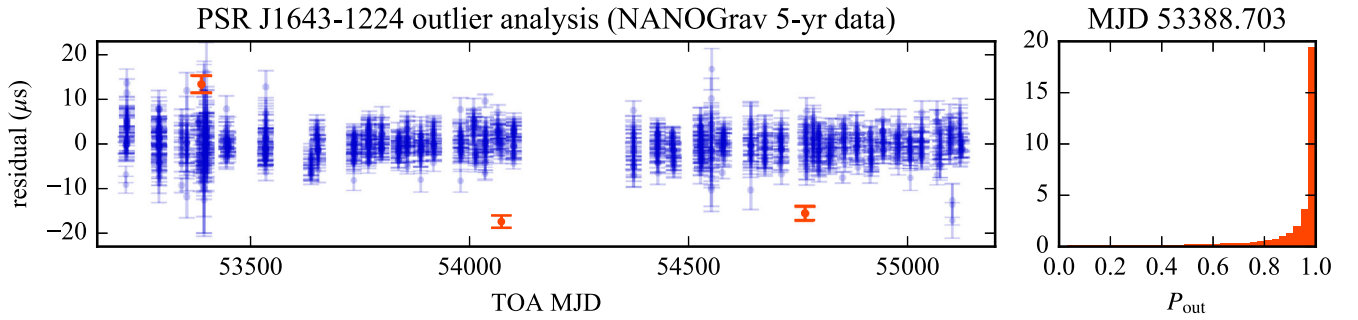
Recent work (also reviewed in van Haasteren & Vallisneri 2014) showed that equation (7) for  $p_{\text{GP}}(\mathbf{y}|\boldsymbol{\eta})$  is completely equivalent to the integral of a *hierarchical* likelihood  $p_h(\mathbf{y}|\boldsymbol{\eta}, \mathbf{c})$ :

$$\begin{aligned} p_{\text{GP}}(\mathbf{y}|\boldsymbol{\eta}) &= \int p_h(\mathbf{y}|\boldsymbol{\eta}, \mathbf{c}) d\mathbf{c} \\ &= \int \frac{e^{-\frac{1}{2} \sum_i \left( y_i - \sum_a \phi_a(x_i) c_a - \sum_{\mu} M_{i\mu} \eta_{\mu} \right)^2 / \sigma_i^2}}{\sqrt{(2\pi)^n \prod_i \sigma_i^2}} \\ &\quad \times \frac{e^{-\frac{1}{2} \sum_{ab} c_a (\Phi_{ab})^{-1} c_b}}{\sqrt{(2\pi)^m \det \Phi}} d\mathbf{c} \\ &= \int \left[ \prod_i p_h(y_i|\boldsymbol{\eta}, \mathbf{c}) \right] \times p(\mathbf{c}) d\mathbf{c}; \end{aligned} \quad (8)$$

here the  $m$  basis vectors  $\phi_a(x_i)$  reproduce the correlated-noise covariance matrix as  $K_{ij} = \sum_{ab} \phi_a(x_i) \Phi_{ab} \phi_b(x_j)$ , the  $c_a$  are known as the basis *weights* and the second exponential factor in equation (8) is effectively a Gaussian prior for the weights (Rasmussen & Williams 2006). See van Haasteren & Vallisneri (2015) for a discussion of how the sums over the  $\phi_a$  converge to analytical covariance expressions in the case of pulsar timing. Thus, we recover equation (7) by marginalizing the hierarchical likelihood with respect to the weights, which is also why equation (7) is known as the *marginalized* pulsar-timing likelihood.

We see immediately from equation (8) that  $p_h(\mathbf{y}|\boldsymbol{\eta}, \mathbf{c})$  factorizes with respect to the individual  $y_i$ , so we can change it into a mixture that accounts for the possibility of outliers. Keeping in mind our picture of timing-model outliers as originating from mistaking noise spikes for pulses, we design a slightly different outlier likelihood than equation (5) – we posit that TOA outliers are distributed uniformly across a pulsar spin period  $P_{\text{spin}}$ . Thus, we make an outlier-tolerant version of equation (8) by way of the simple replacement

$$\begin{aligned} p_h(y_j|\boldsymbol{\eta}, \mathbf{c}) &\rightarrow p_h''(y_j|\boldsymbol{\eta}, \mathbf{c}; \sigma_{\text{out}}, P_{\text{out}}) \\ &= (1 - P_{\text{out}}) p_h(y_j|\boldsymbol{\eta}, \mathbf{c}) + P_{\text{out}} \frac{1}{P_{\text{spin}}}. \end{aligned} \quad (9)$$



**Figure 2.** Outlier analysis of NANOGrav’s five-year PSR J1643–1224 data set (NANOGrav 2013). Left-hand panel: timing residuals, as computed against the best-fitting timing-model and noise parameters. The outlier study identifies three strong outliers with  $P_{i,\text{out}} \simeq 1$ . Right-hand panel: posterior distribution of  $P_{i,\text{out}}$  for the outlier near MJD 53388; in this case,  $P_{i,\text{out}}$  changes slightly across the posterior distribution of timing-model and noise parameters.

Before proceeding with our example, we have three important remarks to make about hierarchical likelihoods in pulsar timing. First, in actual use, we never perform the integral over the weights that we wrote in equation (8) (which would get us back to the marginalized likelihood of equation 7), but rather we sample the weights  $c_a$  in stochastic fashion, together with the timing-model parameters  $\eta_\mu$ . So we work with  $p_h(\mathbf{y}|\boldsymbol{\eta}, \mathbf{c})$  [or, indeed,  $p_h''(\mathbf{y}|\boldsymbol{\eta}, \mathbf{c}; \sigma_{\text{out}}, P_{\text{out}})$ ] rather than  $p_{\text{GP}}(\mathbf{y}|\boldsymbol{\eta})$ .

Our second remark is a consequence of the first because the number of sampled parameters increases considerably with the addition of the weights, adopting an efficient stochastic-sampling scheme becomes paramount. Indeed, earlier attempts to use hierarchical likelihoods (Lentati et al. 2013) were hindered by the difficulty of sampling all the weights efficiently, and in particular by *Neal’s funnel problem* (Neal 2003) of sampling each  $c_a$  together with the variance-like hyperparameter  $\rho_a$  that sets its scale (note that these  $\rho_i$  enter equation 8 implicitly through  $\Phi$ ). Recently, in collaboration with J. A. Ellis, we were able to demonstrate a Hamiltonian sampler (Neal 2011) with NUTS integration tuning (No-U-Turn Sampler: Homan & Gelman 2014), optimized for pulsar-timing hierarchical likelihoods by a chain of data-aware coordinate transformations (van Haasteren, Vallisneri & Ellis, in preparation). The transformations come remarkably close to transforming the target distribution into an easily sampled multivariate Gaussian. The sampler is available in the PICCARD code at [github.com/vhaasteren/piccard](https://github.com/vhaasteren/piccard).

Thirdly, unlike equation (7), the hierarchical likelihood involves no inversion of large, dense matrices ( $\Phi$  is inverted, but it is usually small and diagonal), so its evaluation is orders of magnitude faster than the evaluation of equation (7), especially for large modern pulsar-timing data sets. In practice, this bonus is partially offset by the larger number of parameters to sample, and by the algebraic manipulations required to tame the target probability distribution. Nevertheless, the PICCARD NUTS sampler is remarkably efficient for typical timing-model data sets (van Haasteren, Vallisneri & Ellis, in preparation).

Examples: the outlier-tolerant hierarchical likelihood, sampled with the PICCARD NUTS sampler, solves the contamination problem of the data set in Fig. 1: the three outliers are identified as having  $P_{i,\text{out}} \simeq 1$  (equation 6), whereas all other data points have  $P_{i,\text{out}}$  less than 1 per cent. Most important, as shown by the dotted histogram in the right-hand panel of Fig. 1, the posterior distribution of the EFAC noise parameter becomes unbiased, and tracks closely the posterior obtained by excluding outliers altogether.

For this study, we obtained 20 000 approximately independent samples, each of which describes the timing-model parameters RAJ,

DECJ, PMRA, PMDEC, PX, F0 and F1 (Edwards et al. 2006), as well as the noise hyperparameters EFAC (measurement noise multiplier), EQUAD (quadrature-added noise), the amplitude and exponent of power-law timing noise (represented by 20 sine and cosine Fourier bases) and the outlier hyperparameter  $P_{\text{out}}$ .

Our method can be applied without modification to the real data sets used in pulsar-timing-array searches for gravitational waves. In Fig. 2, we show the outlier analysis of NANOGrav’s five-year PSR J1643–1224 data set (NANOGrav 2013), which was completed in  $\sim 1$  h on a recent multicore workstation, again using the PICCARD NUTS sampler. Three outliers are identified clearly, and shown as the thick red dots and errorbars in the left-hand panel:  $P_{i,\text{out}} \simeq 1$  for the TOAs near MJD 54072 and 54765, and slightly less for the TOA near 53388 (viz.,  $P_{i,\text{out}} = 0.98$ , with posterior distribution corresponding to the integrand of equation 6 shown in the right-hand panel). While these outliers were *not* identified as spurious measurements during the production of the five-year NANOGrav data set, our analysis implies positively that they *are*; luckily, they do not significantly affect the estimation of timing-model or noise parameters.

For this study, we obtained 20 000 approximately independent samples, each describing the timing-model parameters RAJ, DECJ, F0, F1, PMRA, PMDEC, PX, PB, A1, XDOT, TASC, EPS1, EPS2 and 40 DMX dispersion-measure parameters (Edwards et al. 2006; NANOGrav 2013), as well as the noise hyperparameters EFAC (measurement noise multiplier), EQUAD (quadrature-added noise), ECORR (jitter-like epoch-correlated noise), the amplitude and exponent of power-law timing noise (represented by 20 sine and cosine Fourier bases) and the outlier hyperparameter  $P_{\text{out}}$ .

### 3 CONCLUSIONS

We have described an end-to-end, practical method to identify outliers in pulsar-timing data sets, and to perform outlier-robust statistical inference of timing-model parameters and noise hyperparameters. The treatment of outliers is fully consistent: it accounts for time-correlated timing noise, and for the variation of estimated residuals across the posterior distribution of the noise hyperparameters.

Our method relies crucially on the hierarchical form of the pulsar-timing likelihood (equation 8), and on the ability to sample it efficiently, which is now possible with a special-purpose Hamiltonian sampler (van Haasteren, Vallisneri & Ellis, in preparation) freely and openly available at [github.com/vhaasteren/piccard](https://github.com/vhaasteren/piccard). The computational cost of a full inference run scales as  $N_{\text{pars}}^{9/4}$ , where  $N_{\text{pars}}$  is the number of sampled parameters. For current NANOGrav data



sets NANOGrav 2015; NANOGrav, in preparation),  $N_{\text{pars}}$  is dominated by the number of jitter-like noise parameters (one per multi-frequency measurement epoch), which scales linearly with the data set's timespan. Nevertheless, even data sets with  $\sim 20\,000$  TOAs are tractable on workstation-class computers.

Thus, we recommend outlier studies, such as performed above, as a standard step in the production of pulsar-timing-array data sets. Even if a small number of outliers within a large data set is often tolerated well by non-robust statistical inference, the follow up of strong outlier candidates may reveal undetected problems in radio observations and TOA generation. Indeed, we performed our outlier study in the preparation of the NANOGrav 11-year data set (NANOGrav, in preparation). An easily adaptable PYTHON script that performs such a study is available at [github.com/vhaasteren/piccard/outliers](https://github.com/vhaasteren/piccard/outliers).

Reacting to a preliminary version of this work, E. Cameron suggests that pulsar-timing outliers could also be identified in non-outlier-robust noise-parameter estimation by evaluating a marginalized estimator of discrepancy for each data point (Cameron 2016). For instance, if we are sampling the hierarchical likelihood of equation (8), we could compute the normalized residuals  $|y_i - \sum_a \phi_a(x_i)c_a - \sum_\mu M_{i\mu}\eta_\mu|/\sigma_i$ , marginalized against the  $p_h(\eta, \mathbf{c}|\mathbf{y})$  posterior [i.e. averaged over all  $(\eta, \mathbf{c})$  samples]. Even if we are sampling the marginalized likelihood of equation (7), we could still marginalize the residuals by obtaining the joint conditional posterior of the  $(\eta, \mathbf{c})$  as a function of the noise hyperparameters (see van Haasteren & Vallisneri 2014). However, these schemes have computational complexity equivalent to or greater than our mixture method, and they are not fully self-consistent, since the presence of very severe or very frequent outliers may bias the noise posteriors.

Our MNRAS referee asks whether more extensive statements can be made on the practical limits of this technique: for instance, what fraction of outliers can be tolerated and how large the outliers must be to be detected. It is difficult to answer such questions without reference to specific data sets and timing/noise models, because all the degrees of freedom of the fit play together into the probabilistic attribution of each TOAs as a regular data point or an outlier. For the simulated data set of Fig. 1, our scheme fails when the number of outliers reaches  $\sim 10$  per cent of all data points: beyond that level, noise-parameter estimates are biased high and no outliers are identified. For the same data set, outliers can be found when the corresponding residuals exceed  $4\,\mu\text{s}$ ; by comparison, regular data point have average residuals and measurement errors  $\sim \mu\text{s}$ . However, this sensitivity is a function of the outlier model, which in this case makes the strong assumption that outlier residuals are distributed uniformly within the  $\sim 5$ -ms pulsar period. A more conservative choice (e.g. that outlier errors are 10 times the nominal radiometer noise) would make it easier to identify them.

The referee asks also whether our technique can be applied following the discovery of a new pulsar to assist in the search for a phase-connected timing solution. If indeed outliers are biasing initial timing-model fits, then the search should benefit from using an outlier-robust likelihood, such as described in equations (2)–(5). Because noise modelling is not relevant at that stage of the analysis, neither hierarchical likelihoods nor sophisticated sampling would be necessary.

Our work may be extended in multiple directions. The capability of sampling the hierarchical likelihood efficiently (van Haasteren, Vallisneri & Ellis, in preparation) opens up the possibility of a

number of other investigations, such as the characterization of non-Gaussianity (beyond outliers) in timing measurements, similar to what is done by Lentati et al. (2014). A mixture probability (equations 5 and 9) may also be inserted in other places within the probabilistic model of timing noise. For instance, by modifying the prior for the red-noise weights in equation (8) (which has structure  $\propto \exp\{-\mathbf{c}^T \Phi^{-1} \mathbf{c}/2\}$ , with diagonal  $\Phi$ ) one would provide robustness against quasi-monochromatic noise features that may bias the estimation of power-law noise. A similar trick is proposed by Littenberg & Cornish (2010).

## ACKNOWLEDGEMENTS

We thank B. Bassett, E. Cameron, N. Cornish, C. Cutler, J. Ellis, J. Lazio, C. Mingarelli, D. Nice, S. Taylor and the anonymous MNRAS referee for useful comments. MV was supported by the Jet Propulsion Laboratory RTD programme. RvH was supported by NASA Einstein Fellowship grant PF3-140116. This research was supported in part by National Science Foundation Physics Frontier Center award no. 1430284, and by grant PHYS-1066293 and the hospitality of the Aspen Center for Physics. This work was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under contract to the National Aeronautics and Space Administration. Copyright 2016 California Institute of Technology. Government sponsorship acknowledged.

## REFERENCES

- Abdo A. A. et al., 2013, *ApJS*, 208, 17
- Barnett V., Lewis T., 1994, *Outliers in Statistical Data*. Wiley, New York
- Burke-Spolaor S., 2015, *PASP*, preprint ([arXiv:1511.07869](https://arxiv.org/abs/1511.07869))
- Cameron E., 2016, *Fourier Features for Pulsar Timing Models* (Available at: <https://astrostatistics.wordpress.com/2016/09/12/fourier-features-for-pulsar-timing-models/>)
- Cordes J. M., Lazio T. J. W., 2002, preprint ([astro-ph/0207156](https://arxiv.org/abs/astro-ph/0207156))
- Cordes J. M., Lazio T. J. W., 2003, preprint ([astro-ph/0301598](https://arxiv.org/abs/astro-ph/0301598))
- Edwards R. T., Hobbs G. B., Manchester R. N., 2006, *MNRAS*, 372, 1549
- EPTA, 2015, *MNRAS*, 453, 2576
- Hawkins D., 2013, *Identification of Outliers*. Springer Science & Business Media, Berlin
- Hlozek R. et al., 2012, *ApJ*, 752, 79
- Hogg D. W., Bovy J., Lang D., 2010, preprint ([arXiv:1008.4686](https://arxiv.org/abs/1008.4686))
- Homan M. D., Gelman A., 2014, *J. Mach. Learn. Res.*, 15, 1593
- Jaynes E., Bretthorst G., 2003, *Probability Theory: The Logic of Science*. Cambridge Univ. Press, Cambridge
- Knight M., Bassett B. A., Varughese M., Hlozek R., Kunz M., Smith M., Newling J., 2013, *JCAP*, 1, 039
- Kunz M., Bassett B. A., Hlozek R. A., 2007, *Phys. Rev. D*, 75, 103508
- Lattimer J. M., Prakash M., 2007, *Phys. Rep.*, 442, 109
- Lentati L., Alexander P., Hobson M. P., Taylor S., Gair J., Balan S. T., van Haasteren R., 2013, *Phys. Rev. D*, 87, 104021
- Lentati L., Hobson M. P., Alexander P., 2014, *MNRAS*, 444, 3863
- Leroy A. M., Rousseeuw P. J., 1987, *Robust Regression and Outlier Detection*. Wiley, New York
- Littenberg T. B., Cornish N. J., 2010, *Phys. Rev. D*, 82, 103007
- Liu J., 2013, *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag, New York
- Lommen A. N., 2015, *Rep. Prog. Phys.*, 78, 124901
- Lommen A. N., Demorest P., 2013, *Class. Quantum Gravity*, 30, 224001
- Lorimer D. R., Kramer M., 2012, *Handbook of Pulsar Astronomy*. Cambridge Univ. Press, Cambridge
- NANOGrav, 2013, *ApJ*, 762, 94
- NANOGrav, 2015, *ApJ*, 813, 65
- NANOGrav, 2016, *ApJ*, 821, 13

- Neal R. M., 2003, *Ann. Stat.*, 31, 705
- Neal R. M., 2011, in Brooks S., Gelman A., Jones G. L., Meng X.-L., eds, *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC, Boca Raton, FL, p. 113
- PPTA, 2015, *Science*, 349, 1522
- Press W. H., 1997, in Bahcall J. N., Ostriker J. P., eds, *Unsolved Problems in Astrophysics*. Princeton Univ. Press, Princeton, NJ, p. 49
- Rasmussen C., Williams C., 2006, *Gaussian Processes for Machine Learning*. MIT press, Cambridge, MA
- Stairs I. H., 2003, *Living Rev. Relativ.*, 6, 5
- van Haasteren R., Vallisneri M., 2014, *Phys. Rev. D*, 90, 104012
- van Haasteren R., Vallisneri M., 2015, *MNRAS*, 446, 1170
- Wolszczan A., Frail D. A., 1992, *Nature*, 355, 145

This paper has been typeset from a  $\text{\LaTeX}$  file prepared by the author.